# METHOD OF FABRICATING A MOS TRANSISTOR WITH LOW GATE DEPLETION





## BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to a method of fabricating a metal-oxide semiconductor (MOS) transistor, and more specifically, to a method of fabricating a MOS transistor with low gate depletion and high gate drivability in nitride read only memory (NROM).

2. Description of the prior art

Nitride read only memory (NROM) comprises a plurality of memory cells and is used to store data. Each memory cell comprises a control gate and a gate dielectric layer of oxide-nitrogen-oxide (ONO) structure. Since the silicon nitride layer of the ONO gate dielectric layer is highly compact, hot electrons tunneling through the MOS transistor become trapped in the silicon nitride layer, which is used as a floating gate for storing data.

Please refer to Fig.1 through Fig.6. Fig.1 through Fig.6 are cross-sectional diagrams of a method of forming an NROM, which comprises both memory arrays and peripheral devices, according to the prior art. As shown in Fig.1, the prior art method provides a semiconductor wafer 10 with a memory array area 11 and a peripheral circuit region 13 defined on the surface of the silicon substrate 12 of the semiconductor wafer 10.

A first step of the prior method is to perform a conventional oxide-nitride-oxide (ONO) process to form an ONO dielectric

layer composed of a bottom oxide layer 14, a silicon nitride layer 16 and a top oxide layer 18 on the surface of the silicon substrate 12. The bottom oxide layer 14 is a silicon oxide layer grown over silicon substrate 12 typically to a thickness of between 50Å and 150Å in a thermal oxidation operation. A typical oxidation temperature is between 750°C and 1000°C. The thickness of the silicon nitride layer 16 is between 20Å and 150Å. The top oxide layer 18 is an oxidative silicon nitride layer or a deposited silicon oxide layer with a thickness between 50Å and 150Å.

The next step, as shown in Fig. 2, involves forming a patterned photoresist layer 20 on the memory array area 11 for defining positions of bit lines 22. The photoresist layer 20 is used as a mask for performing an anisotropic dry etching process to remove the top oxide layer 18 and the silicon nitride layer 16 not covered by the photoresist layer 20 until the surface of the bottom oxide layer 14 or the silicon substrate 12. Following that, an ion implantation process is performed to form a plurality of doped areas in the silicon substrate 12 that function as bit lines 22. The dosage of the ion implantation process is 2-4 x $10^{15}/cm^2$, and the implanting energy is about 50KeV.

As shown in Fig. 3, the photoresist layer 20 is then removed and a thermal oxidation method with a temperature of 800°C ~950°C is used to form an oxide layer 24 with a thickness of 500Å on a top surface of the bit lines 22, separating each ONO dielectric layer. As well, the thermal oxidation method also activates dopants in the doped areas.

On the peripheral circuit region 13 of the semiconductor

wafer 10, the formation of MOS transistors first employs a photo mask comprising patterns of both the memory array area 11 and the peripheral circuit region 13, as shown in Fig.4. The photo mask is first used to perform an etching process on the peripheral circuit region 13 for the purpose of removing the ONO dielectric layer previously formed on the surface of the silicon substrate 12. Thereafter, an oxidation process is executed to form a gate oxide layer 26. To describe this process in greater detail, the photo mask is first used as a photoresist layer (not shown) to shield the memory array area 11. The ONO dielectric layer on the peripheral circuit region 13 is then used as a sacrificial layer for performing an ion implantation process to adjust threshold voltage of MOS transistors on the peripheral circuit region 13. Thereafter, a dry etching process is performed to remove the top oxide layer 18 and the silicon nitride layer 16, and a wet etching process is performed to remove the bottom oxide layer 14. Finally, the photoresist layer is removed and a thermal oxidation process is performed to form a silicon oxide layer with a thickness of 100Å-150Å on the surface of the silicon substrate 12, which functions as a gate oxide layer 26 of the MOS transistor in the peripheral circuit region 13. Due to the presence of the silicon nitride layer 16 in the memory array area 11, the thermal oxidation process does not significantly affect the thickness of the top oxide layer 18.

Following the gate oxide layer 26 growth step, a polysilicon layer is deposited over the semiconductor wafer 10. As shown in Fig.5, the polysilicon layer will create word lines 28 in the memory array area 11. On the peripheral circuit region 13, the polysilicon layer is etched into the form of gate conductive layers 30 for the MOS transistors. Thereafter, a

standard process is performed to complete the formation of the MOS transistors in the peripheral circuit region 13. First, an ion implantation process is performed on the peripheral circuit region 13 to form a lightly doped drain (LDD) 32 for each MOS transistor. Next, a spacer 33 is formed along the side wall of each MOS transistor. After that, another ion implantation process is performed to create a source 35 and a drain 37 in the silicon substrate 12 of both sides of each MOS transistor. Finally, a silicide layer 38 is formed on the top surface of the gate conductive layer 30, as shown in Fig.6. During the ion implantation process that forms the source 35 and the drain 37 of each MOS transistor, the polysilicon layer that comprises the gate conductive layer 30 is also implanted with ions and becomes a doped polysilicon layer. A sequential annealing process is then used to diffuse dopants uniformly in the polysilicon layer.

The current MOS transistors in the peripheral circuit region 13 of NROM use cobalt (Co), titanium (Ti) or molybdenum (Mo) as metals for forming a silicide layer 38. The disadvantage of using these metals is that they consume a great deal of silicon atoms in the gate conductive layer 30. When too much silicon is consumed in the formation of the silicide layer 38, the dopants cannot be uniformly distributed in the gate conductive layer 30, and a portion of the gate is depleted of dopants, resulting in the gate depletion effect. This induces a signal delay in the gate and results in a substantial degradation in device performance.

## SUMMARY OF THE INVENTION

It is therefore a primary objective of the claimed invention

to provide a method of fabricating a metal-oxide semiconductor (MOS) transistor with low gate depletion and high gate drivability in nitride read only memory (NROM).

The claimed invention first provides a semiconductor wafer with a memory array area and a peripheral circuit region defined on the surface of the semiconductor wafer. Then a gate composed of a silicon oxide layer, an amorphous silicon layer, and a silicon germanium layer is formed on the surface of the peripheral circuit region. Following that a spacer, a source and a drain of the MOS transistor are formed around the gate. Finally, a nickel (Ni) layer is formed on the top surface of the gate, and a rapid thermal annealing process (RTA process) with a temperature between 400℃ and 500℃ is performed to form a silicon nickel layer on the top surface of the gate.

In contrast to the prior MOS transistor of a peripheral circuit region, the MOS transistor manufactured by the claimed invention uses a nickel (Ni) layer as the metal material for forming a silicide layer. The silicon nickel layer formed by reacting nickel with the polysilicon top surface of the gate of the MOS transistor will consume a relatively small amount of silicon atoms. The MOS transistor manufactured by the claimed invention uses an amorphous silicon layer and a germanium silicon layer stacked as a gate in order to increase active dopant concentration in the gate. Therefore, the MOS transistor effectively avoids gate depletion effects and greatly improves gate drivability.

These and other objectives of the claimed invention will no doubt become obvious to those of ordinary skill in the art after having read the following detailed description of the

preferred embodiment, which is illustrated in the various figures and drawings.

## BRIEF DESCRIPTION OF THE DRAWINGS

Fig.1 through Fig.6 are cross-sectional diagrams of a method of forming an NROM according to the prior art.

Fig.7 through Fig.9 are cross-sectional diagrams of a method of forming a MOS transistor in NROM according to the present invention.

## DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Please refer to Fig.7 to Fig.9. Fig.7 to Fig.9 are cross-sectional diagrams of a method of forming a MOS transistor in NROM according to the present invention.

In a preferred embodiment, a semiconductor wafer is firstly provided, which has a memory array area (not shown) and a peripheral circuit region 51 defined on the surface of silicon substrate 52. A plurality of NROM memory cells are formed in the memory array area, and each NROM memory cell comprises a MOS transistor and a silicon nitride layer. However, the MOS transistor of the present invention is formed in the peripheral circuit region 51. The method of manufacturing NROM memory cells in the memory array area is firstly forming a patterned ONO dielectric layer on the surface of the silicon substrate 52. Then a plurality of bit lines and a field oxide layer are formed in the silicon substrate 52. Finally, a threshold voltage level adjustment implant for the MOS transistors in the peripheral circuit region 51 is performed, and the ONO dielectric layer on the surface of the peripheral

circuit region 51 is removed.

As shown in Fig.7, a silicon oxide layer is then formed on the surface of the peripheral circuit region 51 and functions as a gate oxide layer 54 of an NMOS transistor or a PMOS transistor. Thereafter, an amorphous silicon layer is formed on the silicon oxide layer and a silicon germanium layer with the chemical composition of $Si_{1-x}Ge_x$, $x = 0.05 \sim 1.0$ is formed in-situ on the amorphous silicon layer. An etching process is then performed to etch the silicon germanium layer, the amorphous silicon layer, and the silicon oxide layer to form a gate 56 of the MOS transistor on the silicon substrate 52. The silicon germanium layer is formed by performing a chemical vapor deposition (CVD) process aerating silane ($SiH_4$), germane ($GeH_4$), and hydrogen at a temperature ranging between 450℃ and 620℃.

As shown in Fig.8, a first ion implantation process is performed to form a lightly doped drain (LDD) 58 of the MOS transistor. Then a spacer 59 is formed around the gate 56, and a second ion implantation process is performed to form two doped areas on the silicon substrate 52 of two relative sides of the gate 56. Next, a high temperature annealing process is performed to drive in the dopants of the two doped areas, forming a source 60 and a drain 62 of the MOS transistor in the silicon substrate 52.

As shown in Fig.9, a nickel (Ni) layer (not shown) is formed on the top surface of the gate. A rapid thermal annealing process (RTA process) with a temperature of between 400℃ and 500℃ is performed to react the nickel layer with the silicon germanium layer of the top surface of the gate so as to form a silicon

nickel layer 64. Finally, the portion of the nickel layer that does not participate in the reaction is removed, and the formation of the MOS transistor is complete.

The MOS transistor in an NROM manufactured by the present invention uses a nickel (Ni) layer as the metal material for forming a silicide layer on the top surface of the gate. The silicon nickel layer formed by reacting the nickel layer with the silicon germanium layer of the top surface of the gate consumes a relatively small amount of silicon atoms in the silicon germanium layer. For the gate, the MOS transistor stacks an amorphous silicon layer and a germanium silicon layer, and the sequential high temperature annealing process diffuses germanium atoms in the silicon germanium layer into the amorphous silicon layer, transforming the amorphous silicon layer into silicon germanium and increasing active dopant concentration in the gate. In contrast to a prior MOS transistor according to the prior art, a MOS transistor manufactured according to the present invention effectively avoids gate depletion effects and greatly improves gate drivability.

Those skilled in the art will readily observe that numerous modifications and alterations of the device may be made while retaining the teachings of the invention. Accordingly, the above disclosure should be constructed as limited only by the metes and bounds of the appended claims.